Fragmentation des entrepôts de données XML

Hadj Mahboubi et Jérôme Darmont

– – – –

Laboratoire ERIC
Université Lumière Lyon 2
{hadj.mahboubi, jerome.darmont}@eric.univ-lyon2.fr

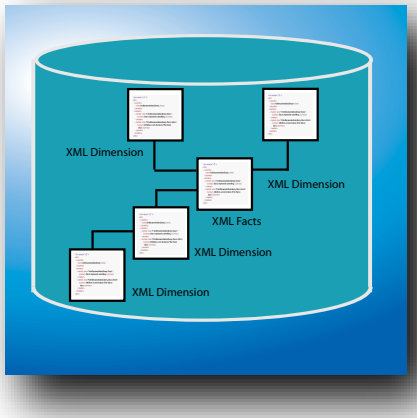
Poitiers, Juin 2007

Plan

Contexte et problématique

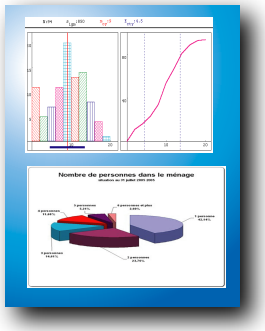
Données XML

Données complexes



Entrepôt de données XML

Analyse



SGBD XML Natif

Implémentation

Volumétrie

Requêtes XQuery coûteuses

Processus d'entreposage et d'analyse de données complexes : Performances ?

Objectifs

Objectifs

Optimiser les performances des requêtes XQuery qui exploitent les entrepôts de données XML : requêtes et volumétrie

Contribution

Adaptation de la fragmentation aux entrepôts de données XML

Fragmentation

diviser un ensemble de données en plusieurs fragments : la combinaison de ces fragments produit l'intégralité des données source, sans perte ou ajout d'information

Fragmentation

Bellatreche (2003):

- fragmentation verticale, horizontale (primaire et dérivée) et hybride

Fragmentation verticale

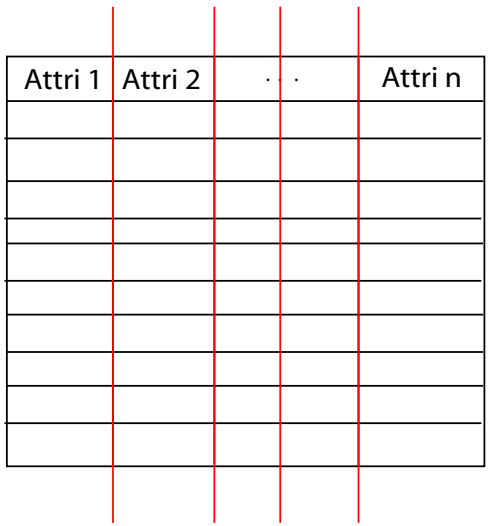
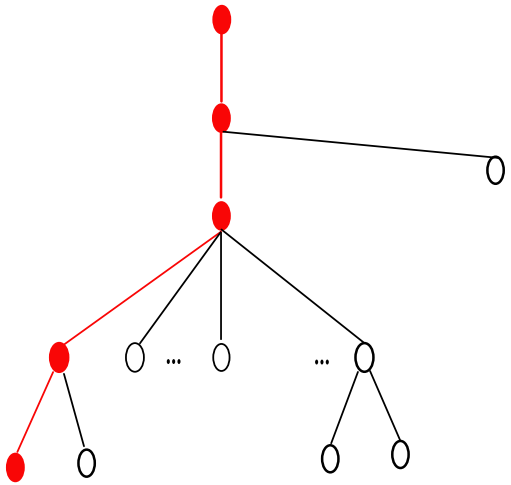


Table / Relation



Document XML

Fragmentation

Bellatreche (2003):

- fragmentation verticale, horizontale (primaire et dérivée) et hybride

Fragmentation horizontale primaire

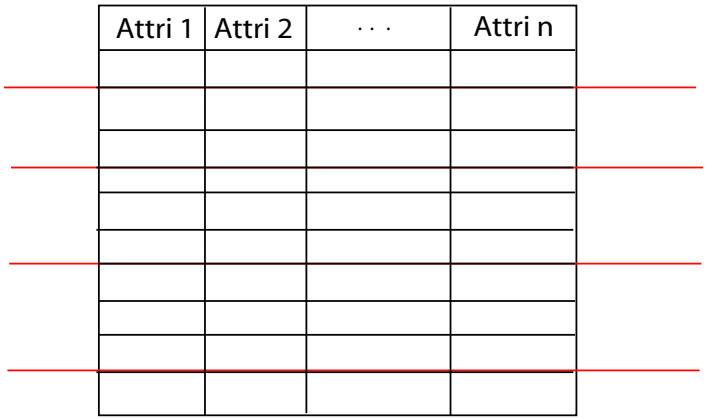
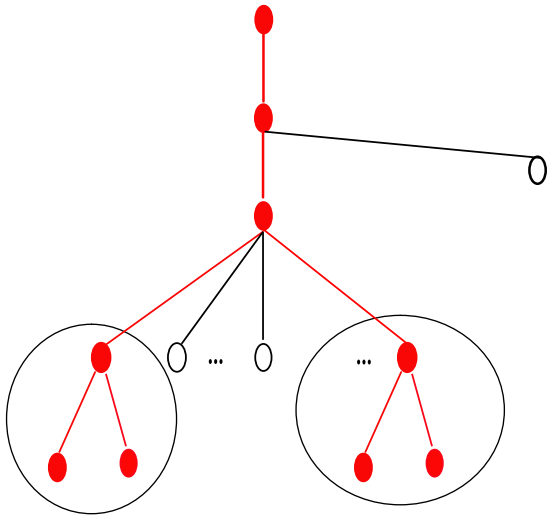


Table / Relation

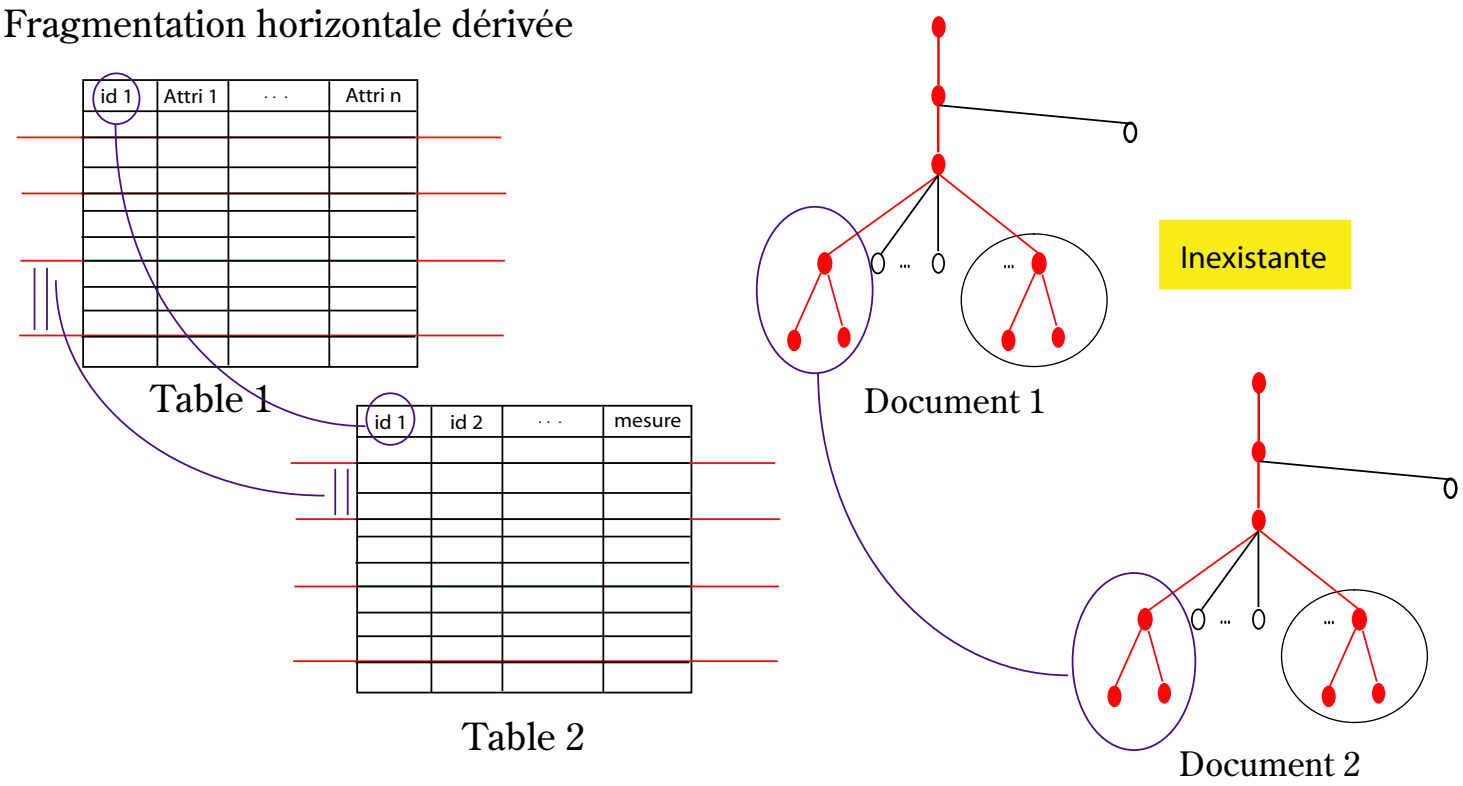


Document XML

Fragmentation

Bellatreche (2003):

- fragmentation verticale, horizontale (primaire et dérivée) et hybride



Etat de l'art

Fragmentation des entrepôts de données

- Contexte : construction d'index Datta et al. (1999), centralisé Noaman et Barker (1999), répartition Bellatreche et Boukhalfa (2005)
- Fragmentation : verticale, horizontale (primaire et dérivée)et hybride
- Contrôle du nombre de fragments : schéma de fragmentation optimal, Bellatreche et Boukhalfa (2005)
- Recommandation de la fragmentation horizontale dérivée pour les entrepôts de données
- Combinaison de la fragmentation horizontale (dérivée) et verticale Wu et Buchmann (1997)

Fragmentation des données XML

- Contexte : répartition Bremer et Gertz (2003), pair-à-pair Bonifati et al. (2004), gestion du cache Bonifati et al. (2006) et collection XML Andrade et al. (2006)
- Fragmentation : verticale, horizontale, hybride et *split* Ma et Schewe (2003)
- Plusieurs manières de définir formellement les fragments (Algèbre XML et expressions de chemin)

Motivation

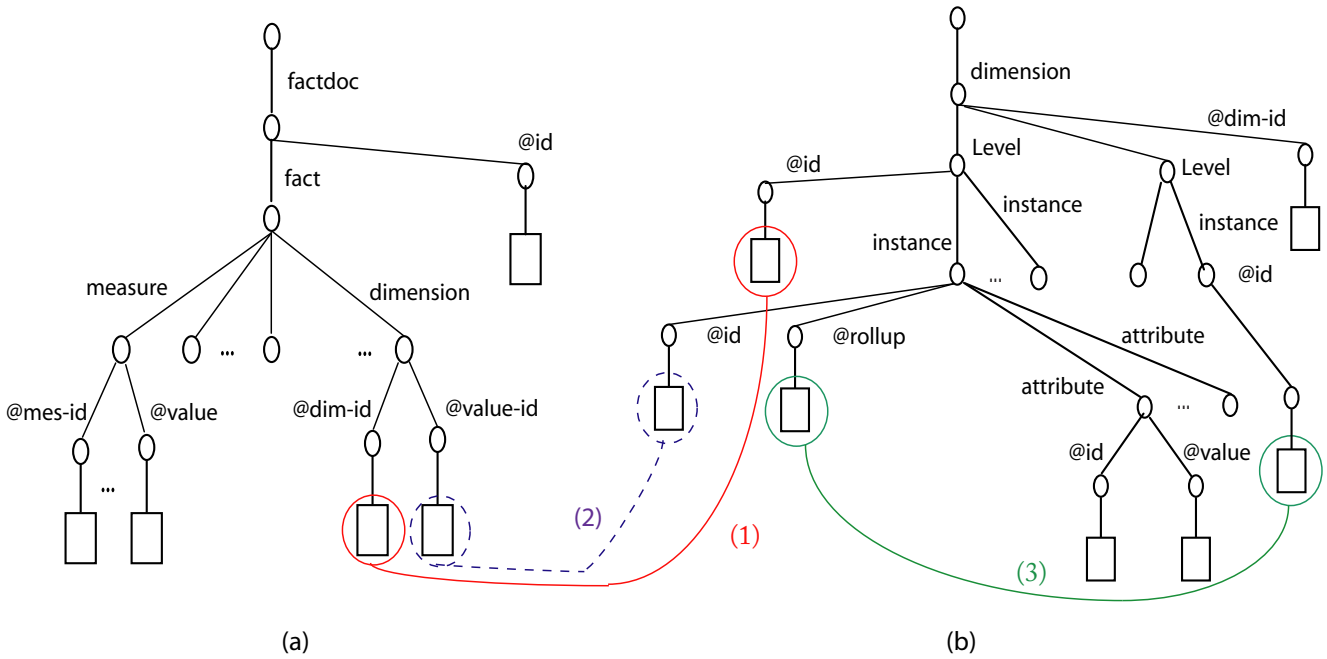
Techniques de fragmentation existantes

- XML : limitation à la fragmentation horizontale primaire (Ma et Schewe,2003; Gertz et Bremer, 2003)
- XML : application de la fragmentation sur un seul document XML
- XML : approches pour les flux de données ne supportent pas la fragmentation horizontale
- ED relationnels : recommandation de la fragmentation horizontale dérivée (Wehrle et al., 2005; Bellatreche et Boukhalfa, 2005)

Fragmentation des entrepôts de données XML

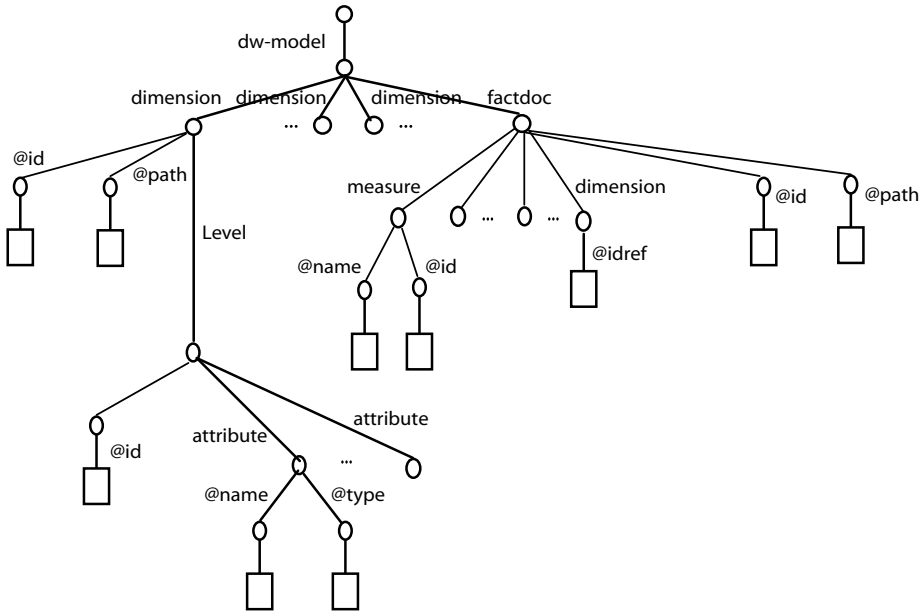
- Adaptation de la fragmentation horizontale dérivée aux entrepôts de données XML
- Diviser les faits en fonction d'un ou de plusieurs fragments des dimensions
- Utile lors du traitement des requêtes de jointure

Entrepôt de données XML : architecture de référence



Graphes des documents *facts.xml* (a) et $dimension_d$*.xml* (b)

Entrepôt de données XML : architecture de référence



Graphe du document *dw-model.xml*

Fragmentation horizontale dérivée : définitions préalables

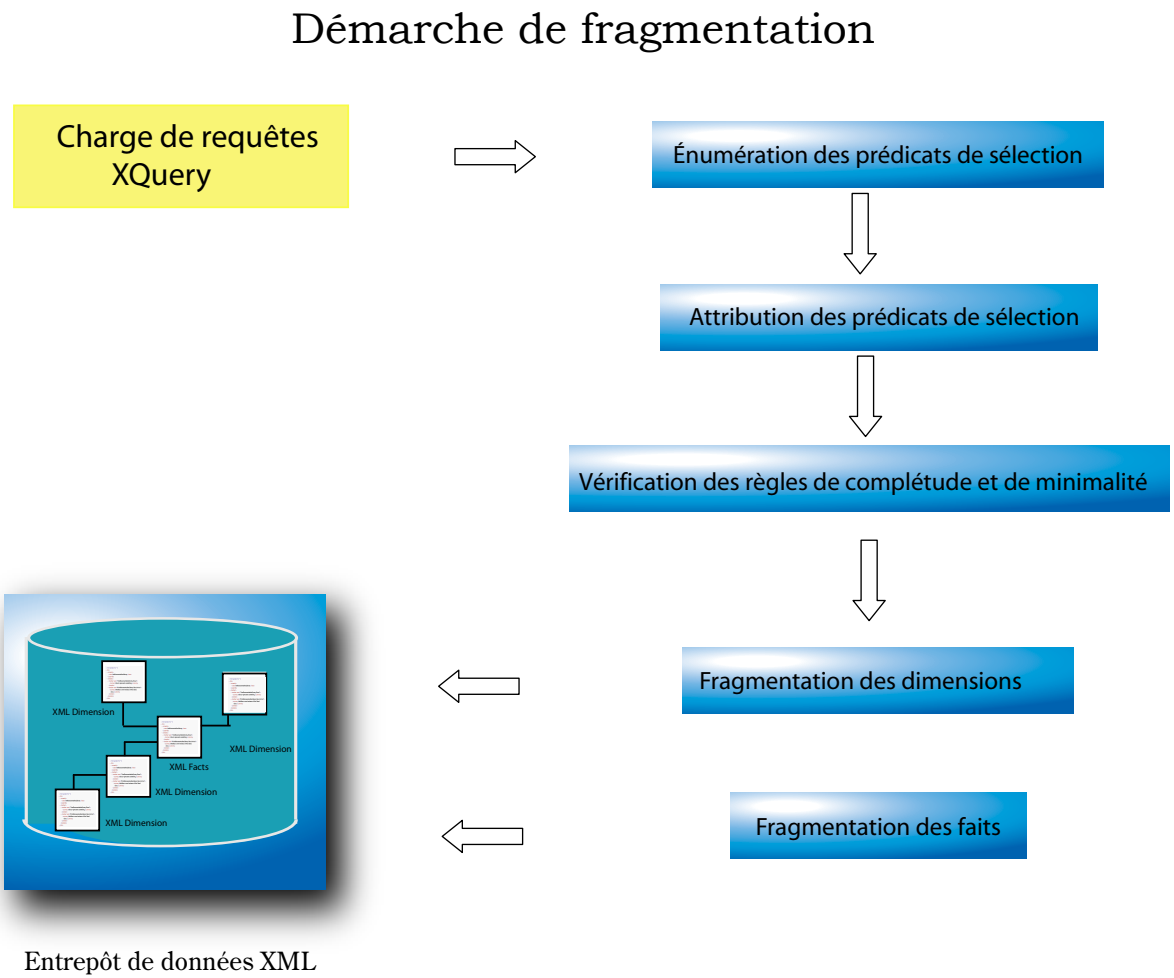
Fragmentation horizontale dérivée : trois informations indispensables

1. **Noms des graphes propriétaires et du graphe membre** : $G_{dimension_d}$ et G_{facts}
2. **Nombre de fragments horizontaux des dimensions** : fragmentation horizontale primaire des $G_{dimension_d}$
3. **Qualification de la jointure** : fragmenter le graphe G_{facts} en fonction des fragments $G_{dimension_d}$

Qualification de la jointure

document(*facts.xml*)/FactDoc/dimension/[@dim-id=document(*dimension$_i$.xml*)
/dimension/Level/@id]
and
document(*facts.xml*)/FactDoc/dimension/[@value-id=document(*dimension$_i$.xml*)
/dimension/Level[@id=@dim-id]/instance/@id]

Démarche de fragmentation



Démarche de fragmentation (1/4)

- **Énumération des prédicats de sélection**
  - Identifier les prédicats de sélection p d'une charge de requêtes
  - Analyse syntaxique de la charge
  - Clause Where :
    $p := [a/attribute/@name =' cust\_city'$
    $and\ a/attribute/@value =' Lyon']$

```
for $a in //dimension/Level[@id='customers']/instance,
$x in //CubeFacts/cube/Cell
...
where $a/attribute/@name='cust_city'
and $a/attribute/@value='Lyon'
and $x/dimension /@node=$a/@id
and $x/dimension/@id='customers'
group by(cust_name,@cust_zip_code)
return name='cust_name', aggregation(sum, quantity)
```

Démarche de fragmentation (2/4)

- **Attribution des prédicats de sélection aux graphes XML des dimensions**
  - Affecter à chaque graphe $G_{dimension_d}$ un ensemble de prédicats, $Exp_{dimension}$
  - Expression de chemin $root_t//dimension/@id =' Nom\ de\ la\ dimension'$
  - Identification des graphes de dimension à fragmenter, $G_{candidat}$

```
for $a in //dimension/Level[@id='customers']/instance,
$x in //CubeFacts/cube/Cell
...
where $a/attribute/@name='cust_city'
and $a/attribute/@value='Lyon'
and $x/dimension /@node=$a/@id
and $x/dimension/@id='customers'
group by(cust_name,@cust_zip_code)
return name='cust_name', aggregation(sum, quantity)
```

Démarche de fragmentation (3/4)

- **Vérification des règles de complétude et de minimalité des prédicats de sélection**
  - Application de l'algorithme des *COM-MIN* (Özsu et Valduriez, 1999; Bellatreche, 2000)
  - S'assurer qu'un graphe candidat est divisé en au moins deux fragments
  - Ensemble complet et minimal de prédicats pertinents pour la fragmentation $Exp'_{dimension}$
  - $p := [a/attribute/@name =' cust\_city' \ and \ a/attribute/@value =' Lyon']$

    p respecte la règle de l'algorithme *COM-MIN*

Démarche de fragmentation (4/4)

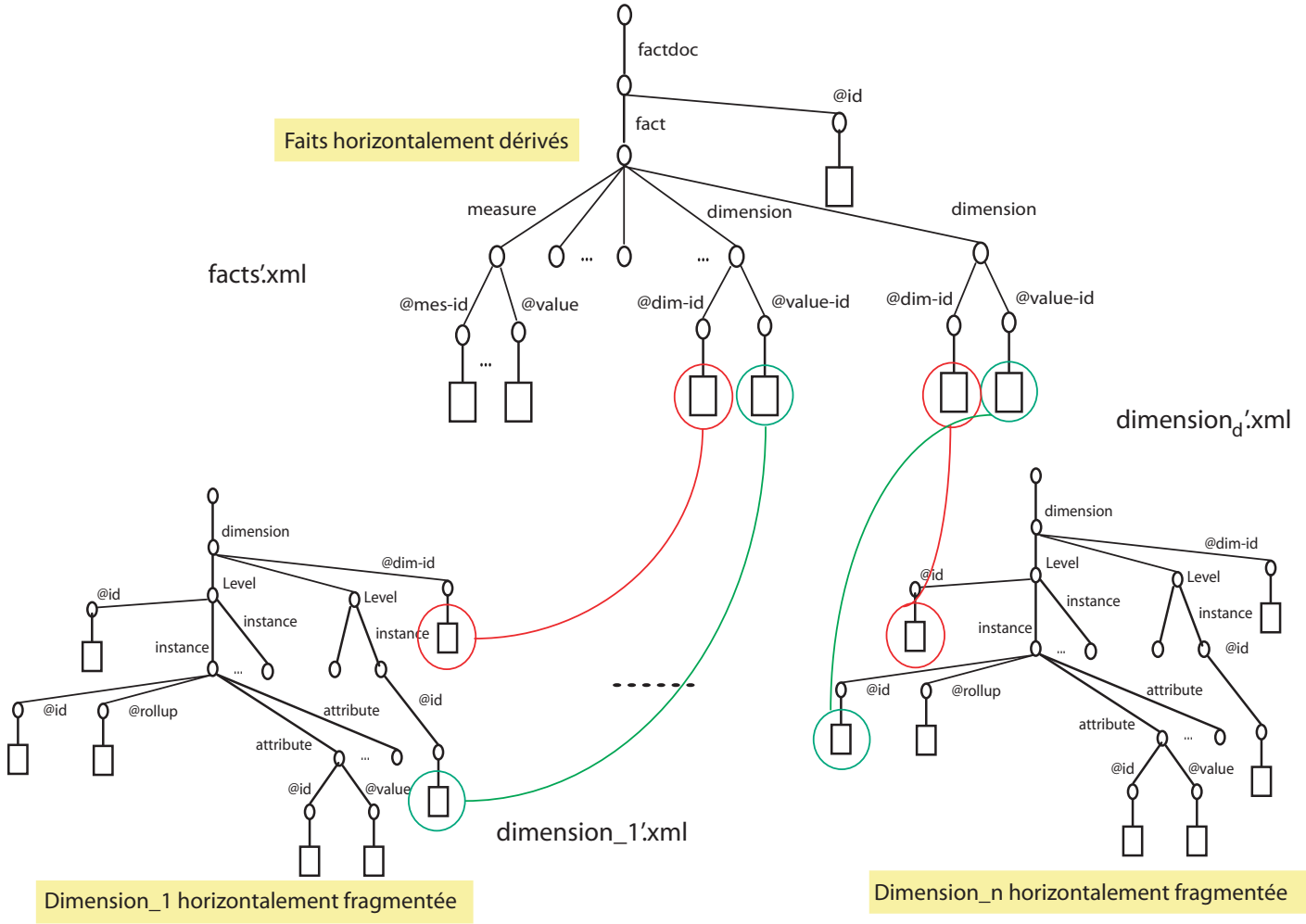
- **Fragmentation des graphes $G_{candidat}$**
  - Fragmentation des graphes de dimension $G_{dimension_d}$: application d'un algorithme de fragmentation primaire, (Özsu et Valduriez, 1999)

- **Fragmentation du graphe G_{facts}** en fonction des fragments définis par le schéma de fragmentation primaire horizontale

  Nombre de fragments (Sous-schémas) $N_{fact} = \prod_{d=1}^{n} N_d$

  - Requêtes XQuery : qualification de la jointure

Exemple de fragment



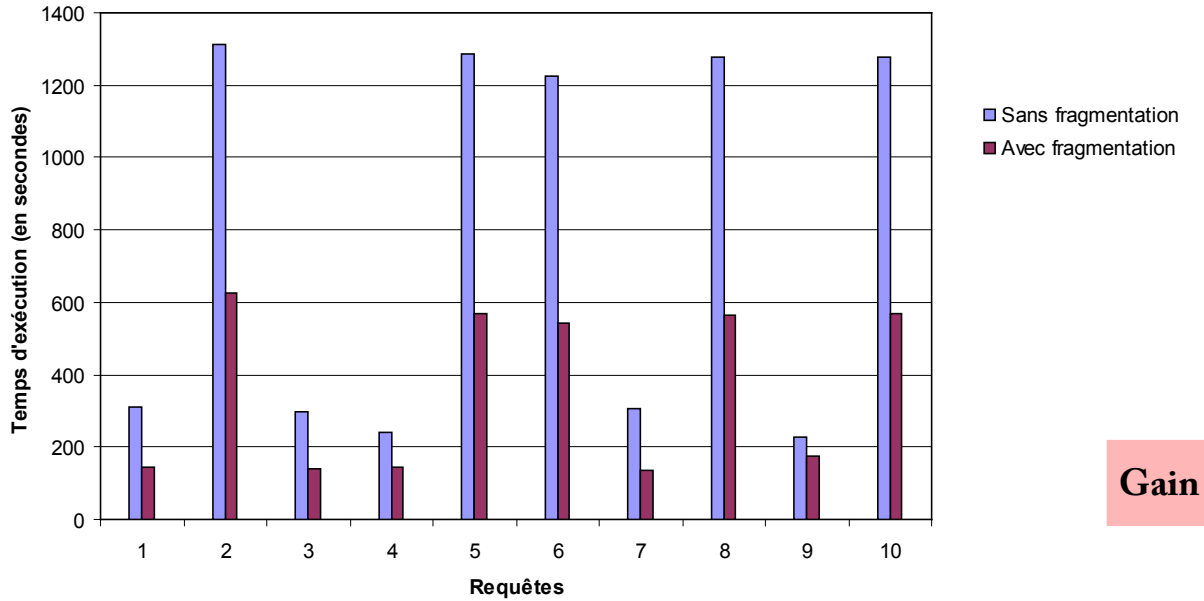
Expérimentations

Entrepôts + charge : Banc d'essai XWB (XML Data Warehouse Benchmark)

4 documents pour les dimensions (*products*, *customers*, *supplies* et *date*)
et 1 document pour les faits (*sales*)

SGBD : X-Hive

16 fragments --> Collections différentes



Gain de 46 %

Temps d'exécution de la charge avec et sans notre démarche de fragmentation

Conclusion et perspectives

Conclusion

1. Adaptation de la fragmentation horizontale dérivée des entrepôts de données XML
2. Premières expérimentations : réduction du temps d'exécution de requêtes décisionnelles XQuery de façon significative avec un gain de 46 %

Perspectives

1. Approfondir les expérimentations
   - Mécanisme de traitement de requêtes distribuées
   - Expérimentation sur différentes configurations d'entrepôts de tailles variées
2. Traiter la problématique du nombre fragments (sous-schémas) générés par une fragmentation horizontale dérivée
3. Distribution des entrepôts de données XML
   - Identifier l'architecture de répartition des fragments XML : architecture de base de données répartie classique, réseau pair-à-pair ou grille de données